Genome Annotation in ERIC

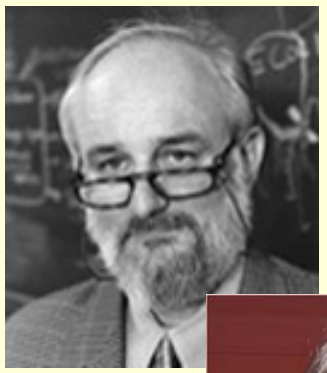
Guy Plunkett III
University of Wisconsin-Madison

BRC2 Meeting
Virginia Tech
May 16, 2005

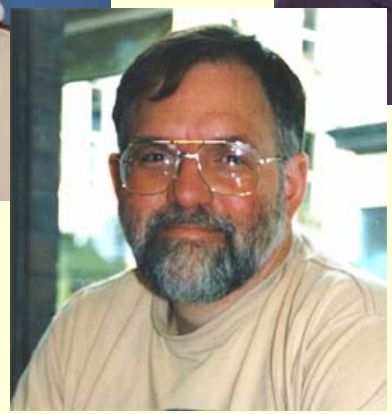
ERIC Team:

ERIC Curators:

- Guy Plunkett III, Ph.D. - Senior Curator;
  David Bowen, Ph.D.; Val Burland, Ph.D.; Eric Cabot, Ph.D.;
  Jeremy Glasner, Ph.D.







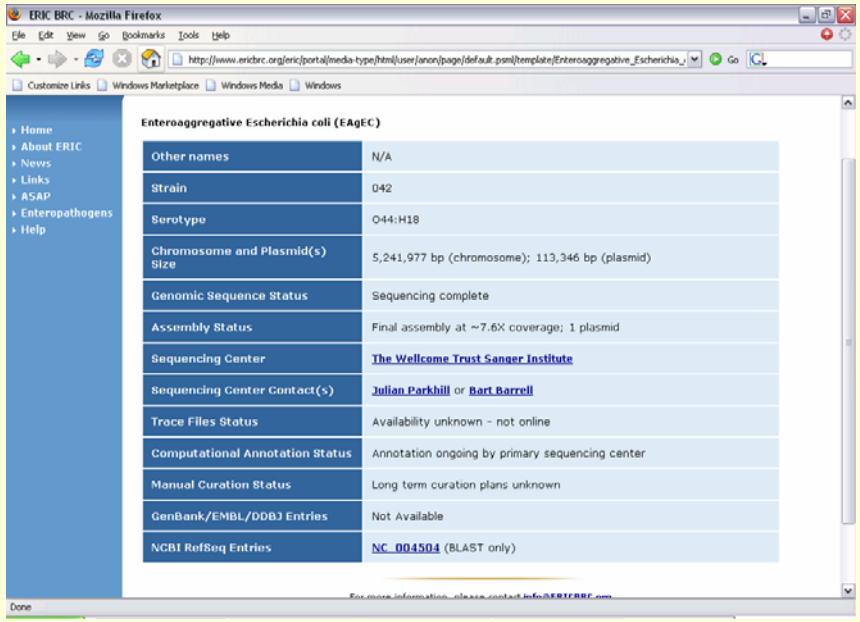

Enteropathogen Resource Integration Center

diarrheagenic *E. coli* genomes published and in progress

- EHEC O157:H7 (2 genomes in ERIC)
- EAgEC (Sanger Centre)
- EPEC (Sanger Centre)

13 others, including

- NIH-MGP (TIGR)
- Universität Münster
- Kitasato University
- Institut Pasteur and Genoscope



**thanks to J. Kaper*

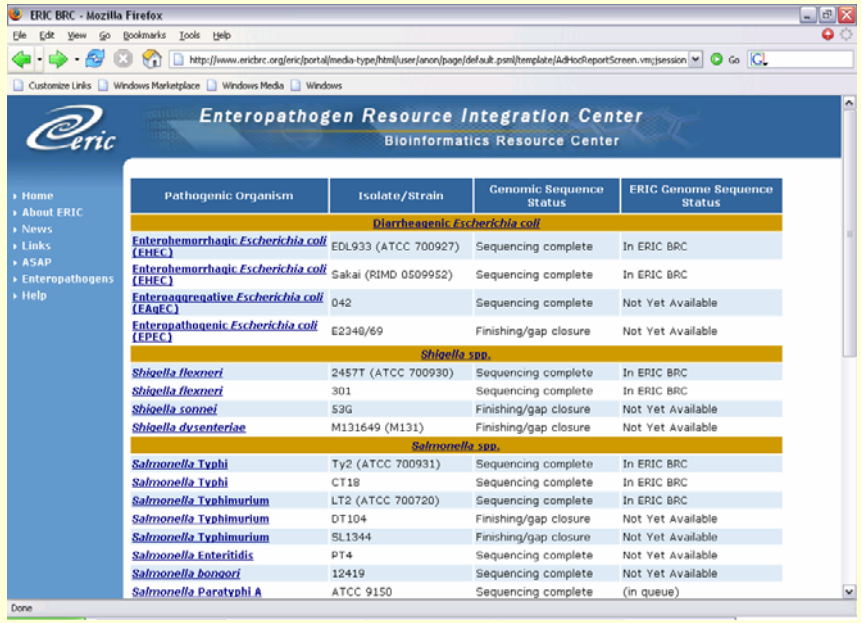
Enteropathogen Resource Integration Center

Shigella genomes published and in progress

- *S. flexneri* (2 genomes in ERIC)
- *S. sonnei* (Sanger Centre)
- *S. dysenteriae* (Sanger Centre)

others, including

- *S. boydii* (TIGR)
- *S. dysenteriae* (TIGR)
- several from China?



Enteropathogen Resource Integration Center

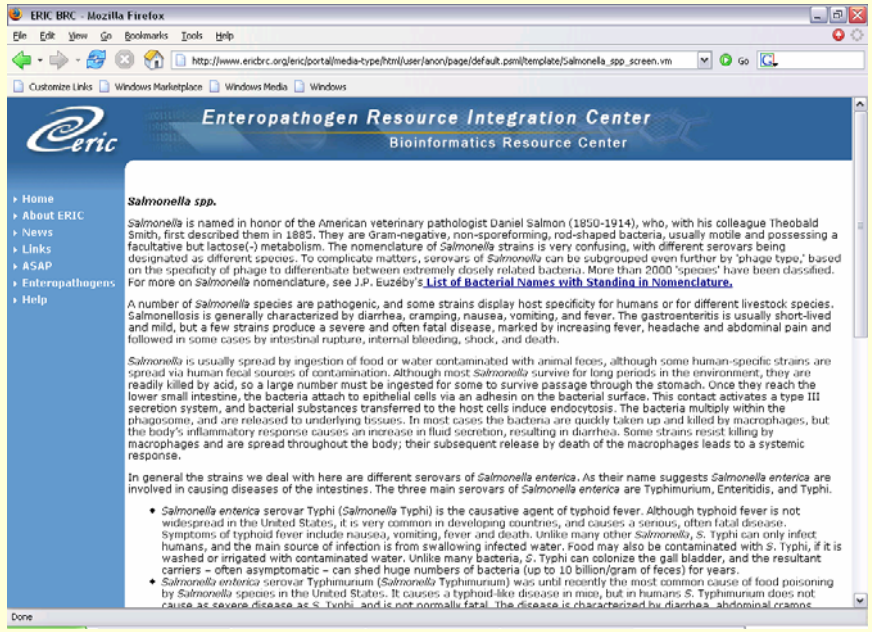
Salmonella genomes published and in progress

- Typhi (2 genomes in ERIC)
- Typhimurium (1 genome in ERIC)
- Paratyphi A (1 genome in ERIC)
- Choleraesuis (1 genome in ERIC)

12 others, including

- Sanger Centre
- Washington University
- University of Illinois

***salmonella.org**



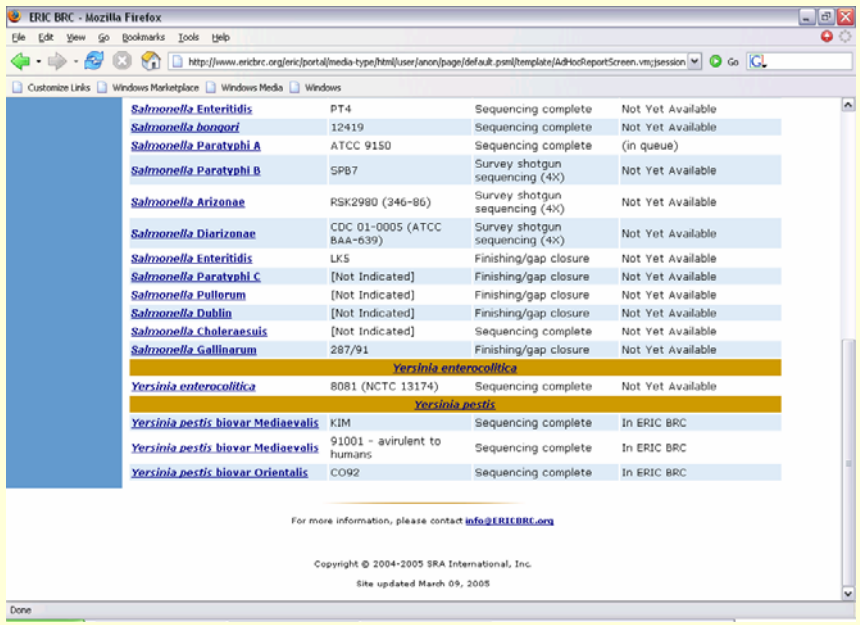
Enteropathogen Resource Integration Center

Yersinia genomes published and in progress

- *Y. pestis* (3 genomes in ERIC)
- *Y. enterocolitica* (Sanger Centre)

10 - 12 others including

- TIGR



Enteropathogen Resource Integration Center

in addition to whole genomes, other relevant sequences that might be added include:

plasmids, bacteriophage, “pathogenicity islands,” and other mobile genetic elements

e.g., pWR501 (*S. flexneri* 5A), R27 (*Salmonella* Typhi)

contigs from genome-scanning projects

Enteropathogen Resource Integration Center

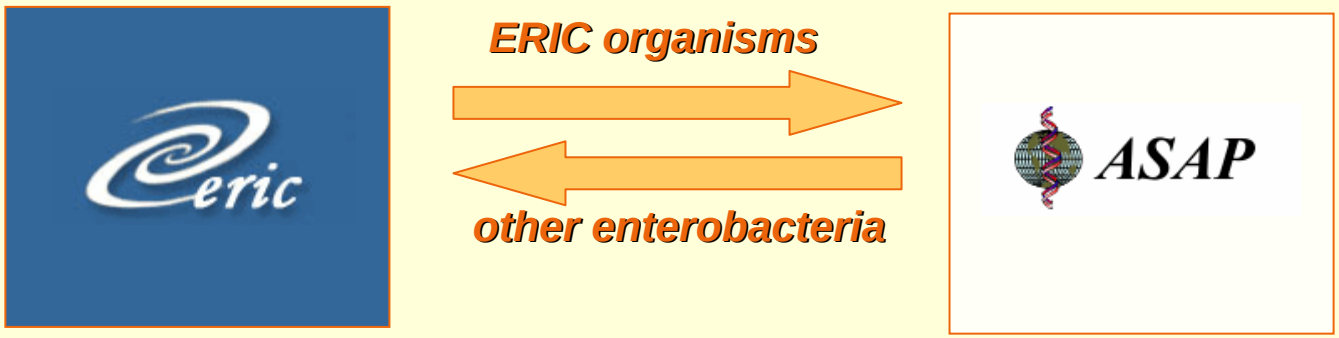
related genomes outside the BRC set are mirrored from UW for comparative purposes:

commensal *E. coli* (K-12, others in progress)

extraintestinal *E. coli* (CFT073, others in progress)

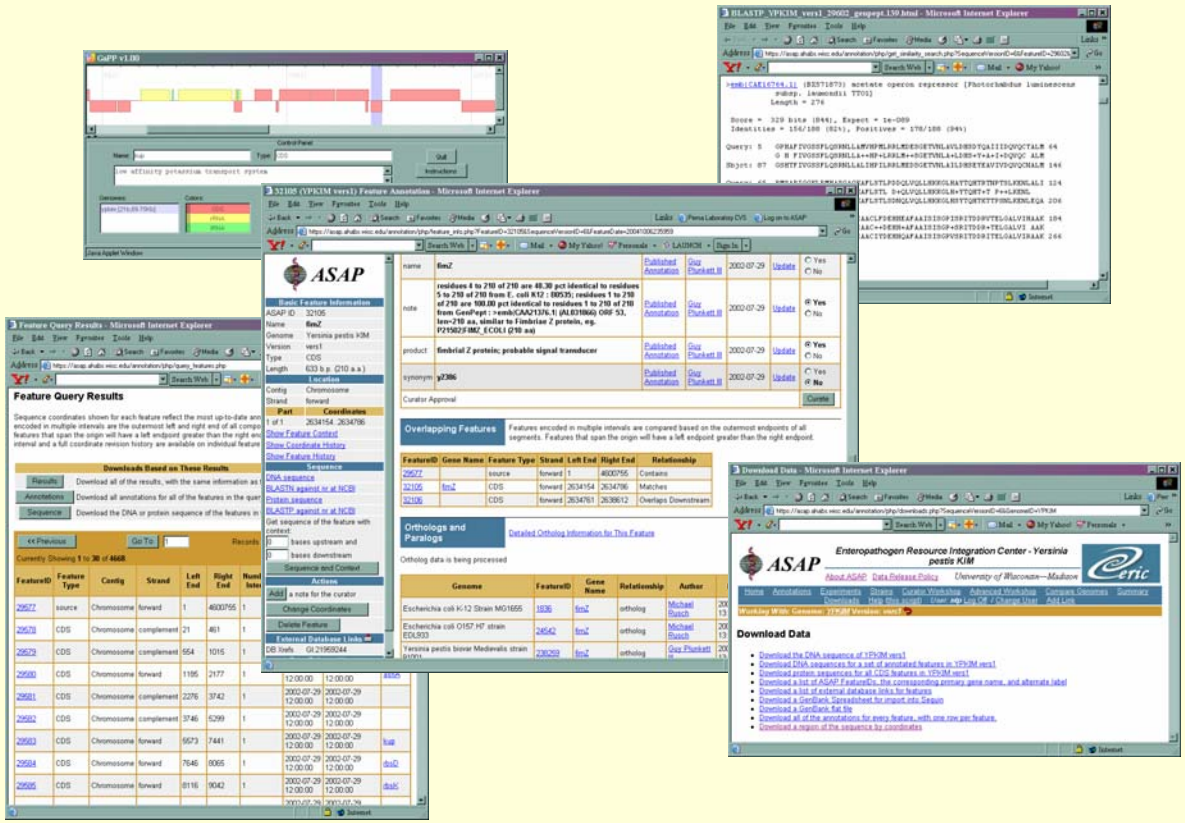
Y. pseudotuberculosis

… and others as they become available



Enteropathogen Resource Integration Center

Bioinformatics Resource Center



ASAP (A Systematic Annotation Package) for community annotation of genomes **meets needs for direct, community annotation (with authorship and annotation history tracking), multiple annotations of features, evidence codes, using controlled vocabularies, curatorial review, support of cross-genome comparisons, and web-based updating and access.**

Enteropathogen Resource Integration Center

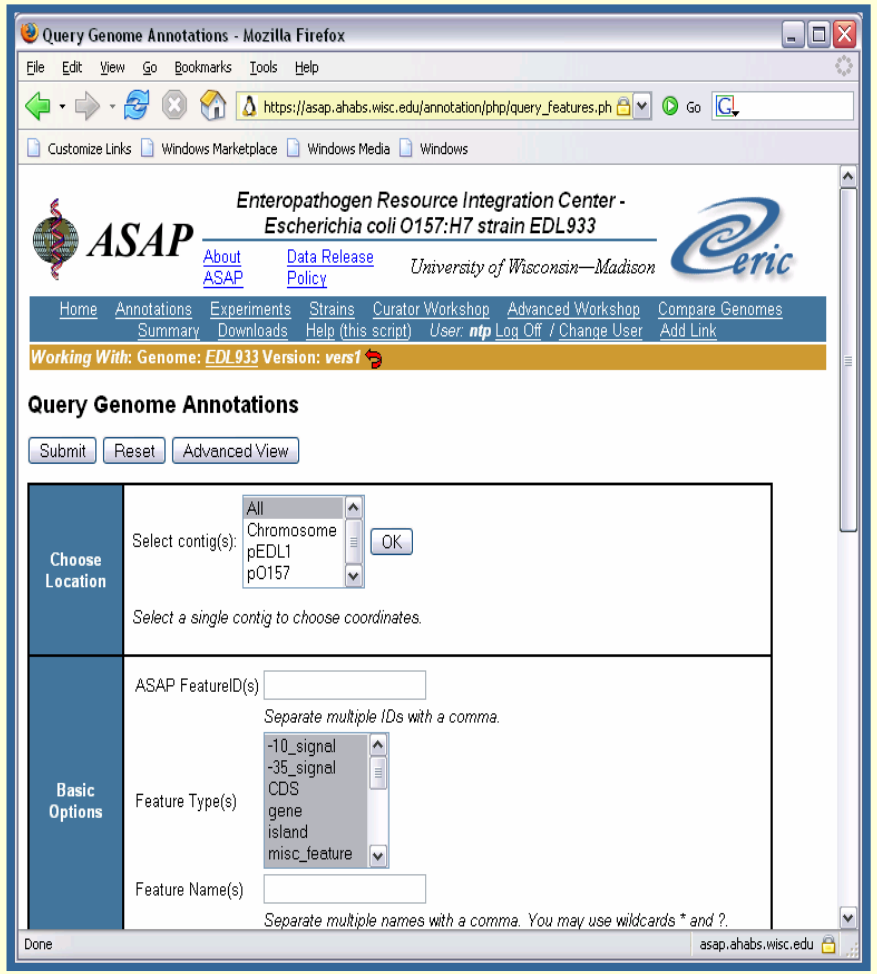
ERIC-BRC Annotation

ASAP can be populated with:

- Unannotated draft genome sequences
- Preliminary annotations in tabular format and associated sequences in FASTA format
- Genbank flatfiles

- For the *E. coli* and *Shigella spp.* being sequenced at TIGR, plan to populate ASAP with annotations from the TIGR automated analysis pipeline

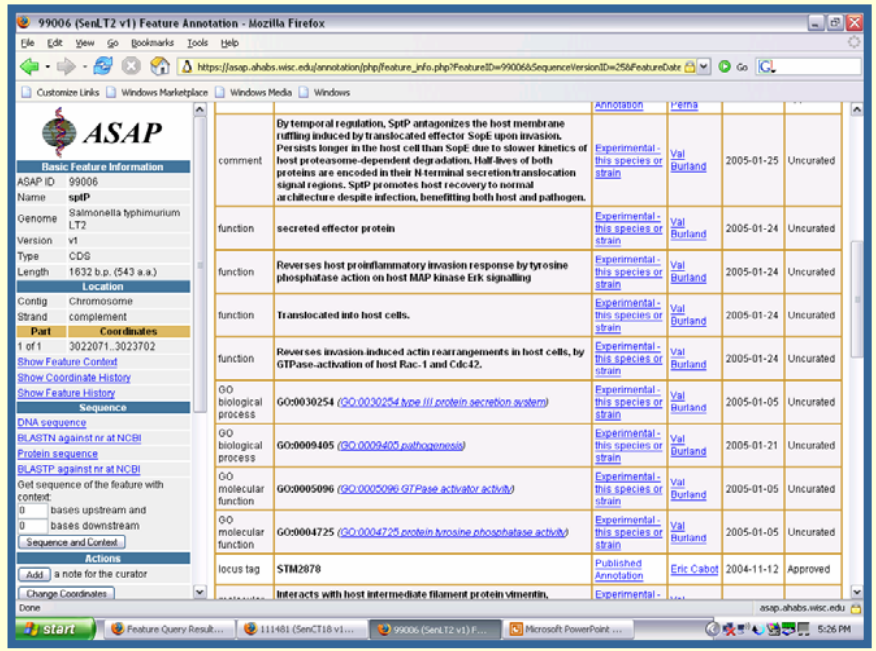
- Can track versions of sequences and versions of annotations



Enteropathogen Resource Integration Center

ERIC-BRC Annotation

- All standard INSD Feature types are supported, plus additional custom types: conserved segment, contig, delimiter (for separating contigs in a “pseudomolecule,” island, prophage
- For each Feature, a detailed annotation page supports standard and custom annotations (aka “qualifiers”)
- Basic information: type of Feature (e.g., CDS, rRNA, tRNA), version of the genome, location within the genome (coordinates and orientation), actual sequences (nt, aa, flanking)
- Rich Annotations: comments, function including Gene Ontology (GO), molecular interaction, mutant phenotypes, structures in PDB



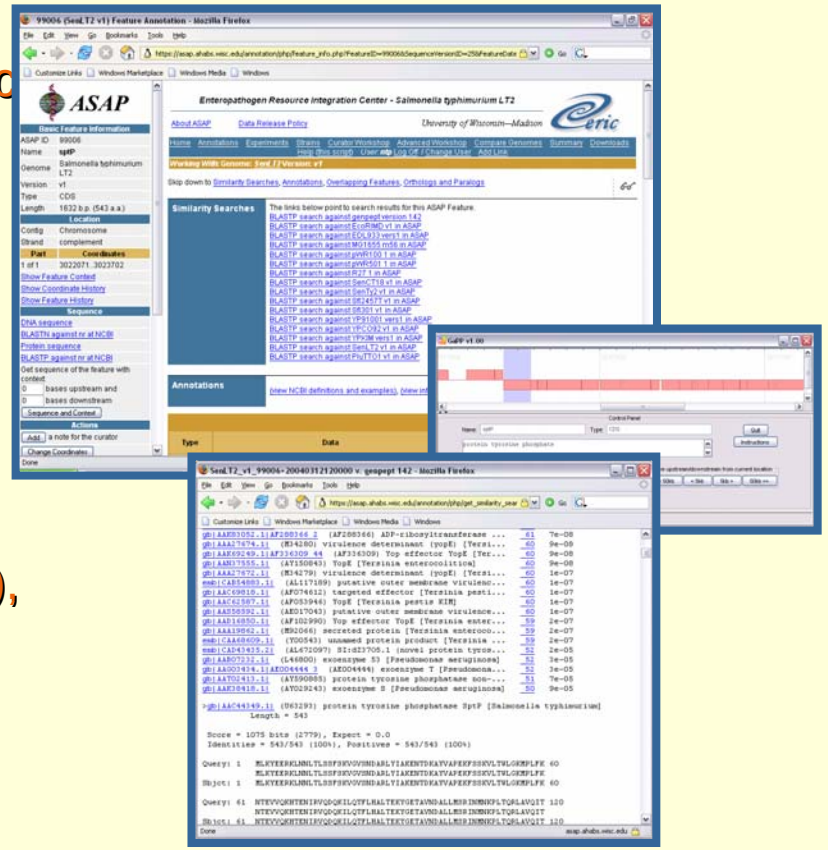
Enteropathogen Resource Integration Center

ERIC-BRC Annotation

- Evidence Code, Reference and Author
- Annotations are immediately visible to public but marked as ‘uncurated’
- Designated ERIC-BRC curators review all annotations

- Current curator-reviewed and community contributed annotations, as well as sequences, can be queried interactively, downloaded in standard formats (Genbank flatfile, tabular text and FASTA, soon GFF3), extracted for deposition in Genbank

- Orthologs, experimental data, BLAST at NCBI and selected pre-run BLAST reports, links to other resources



Enteropathogen Resource Integration Center

ERIC-BRC Annotation

GOALS

- up-to-date core annotations for existing and new genomes
- standardization across genomes where appropriate
- value added through
  - annotation/review by 5 dedicated curators
  - community contributions

STRATEGIES

- leverage similarities across genomes
- focus on genes and processes relevant to BRC goals

Enteropathogen Resource Integration Center

ERIC-BRC Annotation

TASK BASED APPROACH TO ANNOTATION

Tasks are motivated by:

- relevance to virulence/pathogenicity
- community interest
- urgent need to standardize features across genomes

Advantages:

- curators develop and apply expertise in focused areas
- genes with related functions are handled together
- related genes from different genomes are handled together

Enteropathogen Resource Integration Center

ERIC-BRC Annotation

- Attachment and colonization:
  - Fimbriae and adhesins
  - Quorum sensing
  - Tissue tropism
  - Ability to use specific nutrients localized in the target zone
  - Resistance to host defense (acid resistance; urease, serum resistance)
  - Drug resistance mechanisms (efflux pumps)
- Invasion of host cells:
  - Effectors translocated into the host cell
  - Cause changes to host cell morphology
  - Persistence in host cell
  - Resistance to or evasion of host defense mechanisms (oxidation, phagocytosis)
  - Interaction with internal host cell membranes (vacuoles, Golgi)
  - Bacterial motility in host cells
  - Replication in host cells
  - Mechanisms of dissemination to blood and other tissues
  - Suppression of host immune response
  - Regulation of invasion systems: bacterial sensor/effector systems, inducible gene expression
- Secretion systems:
  - Types I – IV mechanisms and components, specialization
  - Chaperones

Enteropathogen Resource Integration Center

ERIC-BRC Annotation

- Toxins:
  - Heat-stable enterotoxins
  - Heat-labile enterotoxins
  - AB-type toxins
  - RTX-type
  - Hemolysins
  - Tc-type
  - Toxin delivery mechanisms
- Serotypes and surface antigens:
  - Colanic acid
  - O-antigen
  - Enterobacterial Common Antigen (ECA)
  - LPS
  - H- and K-antigens
- Islands, phages, plasmids and mobile DNA
- Drug resistance
- Iron acquisition and storage
- Aerobic/Anaerobic pathways
- Fatty acid and polyketide synthesis: multimodular (“factory”) proteins
- The genetic basis for diagnostic tests in current use
- Non-coding small RNAs

Enteropathogen Resource Integration Center

ERIC-BRC Annotation

Start Site Reassessment

- whole-genome proteomics studies and gene-based microarray designs assume accurate assessment of coding sequences
- even the best gene prediction tools have problems with genes acquired by horizontal transfer and those that encode small proteins or proteins with transmembrane domains and/or signal peptides.
- such genes include some of the best candidates for diagnostics (pathogen-specific genes) and vaccine development (*e.g.*, surface-exposed and exported proteins).
- wrong assignments in other genomes are readily propagated to new genomes during annotation.
- identifying coding sequences and annotating the correct translation start, especially in the absence of experimental evidence, relies on a variety of tools, including comparison across ortholog sets, codon usage patterns, RBS (SD) predictions, and even transcription start predictions

Enteropathogen Resource Integration Center

ERIC-BRC Annotation

Pseudogenes

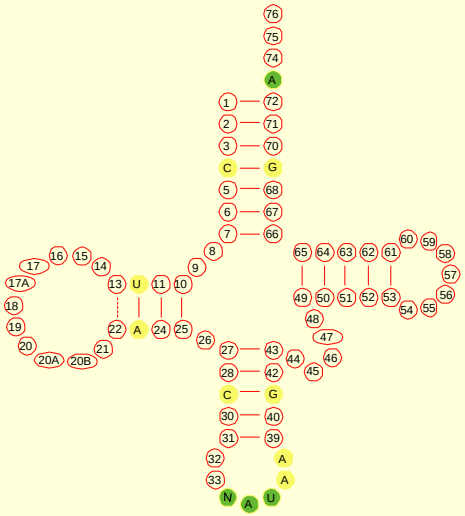
- mutant genes in a genome sequence, usually in comparison to a related genome where the wild-type or “ancestral” state is seen
- distinguished from missense mutations, where the gene is still intact but may have altered functionality
- genes disrupted by in-frame stop codons, frameshifts, insertion of IS elements, prophages, or islands; gene remnants, the aftermath of deletions, rearrangements, etc.
- implicated in the evolution of pathogens and their adaptation to new niches.
- inconsistently annotated in current genomes, and can actually be thought of in two conflicting ways:
  - in the context of evolution and comparative genomics: is a given gene present or absent in a particular genome, and is it intact or disrupted?
  - in the context of a given pathogen, are any partial genes actually translated? is the consequence of a pseudogene something other than the straightforward loss of function?

Enteropathogen Resource Integration Center

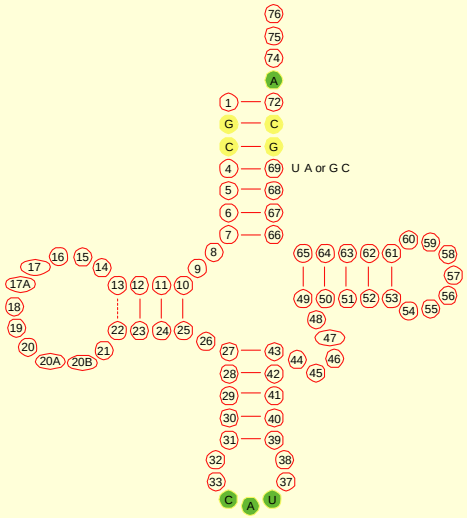
ERIC-BRC Annotation

Fine-tuning tRNA annotations

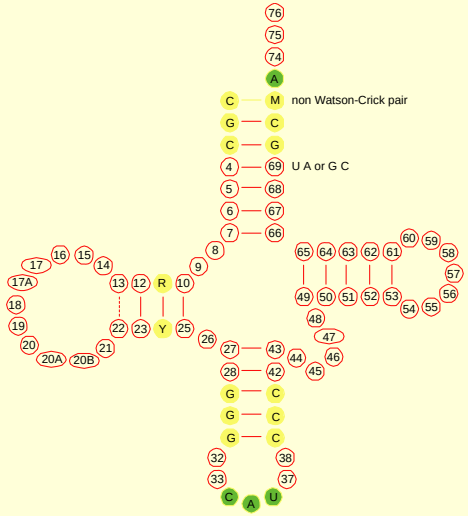
- Islands often integrated into or near tRNA genes
- tRNA genes within islands: improved expression of horizontally acquired genes?
- tRNAscan-SE does an excellent job, but does not distinguish between different tRNAs with CAT anticodons:



tRNA-Ile



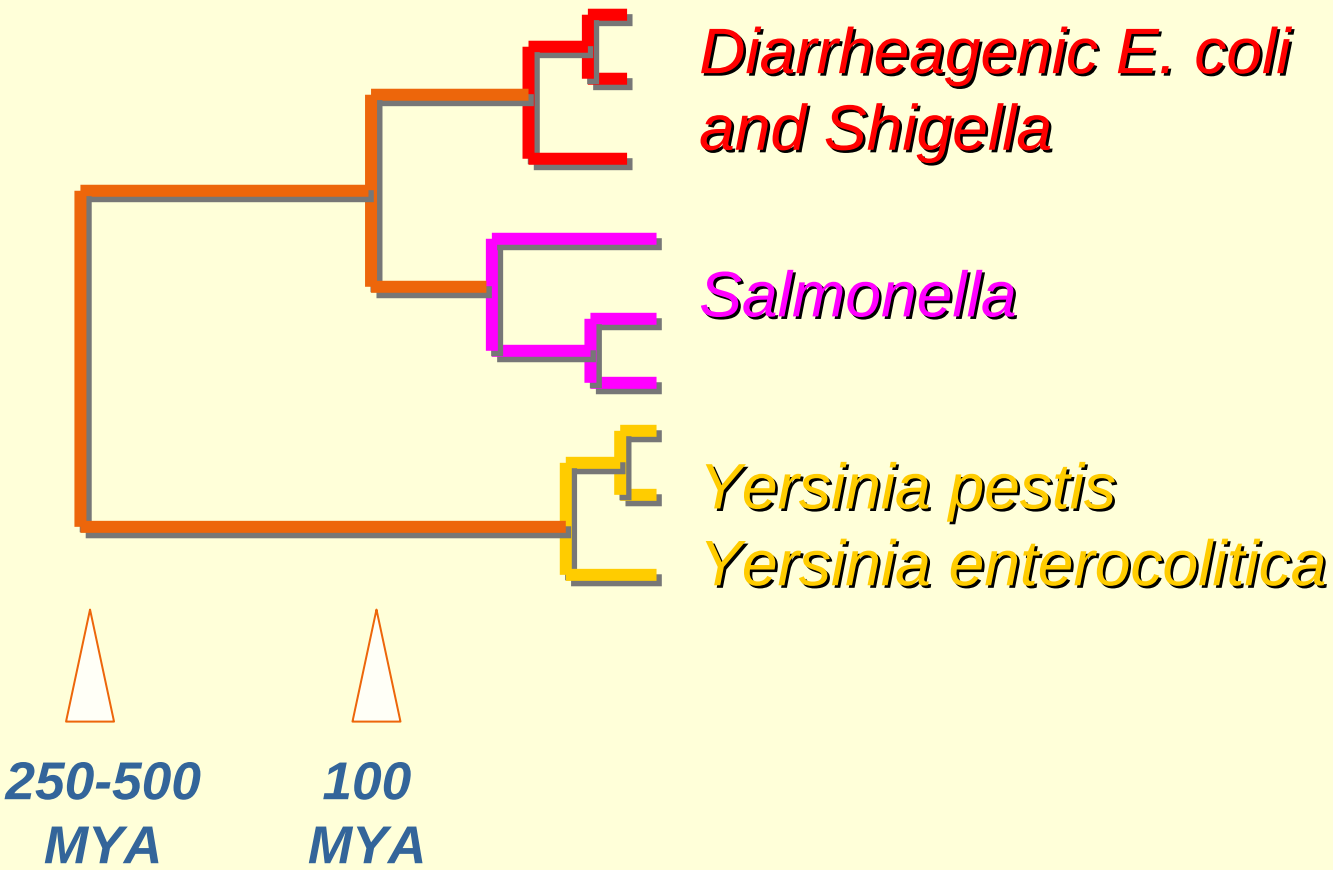
tRNA-Met
(elongator)



tRNA-fMet
(initiator)

Enteropathogen Resource Integration Center

ERIC-BRC Annotation

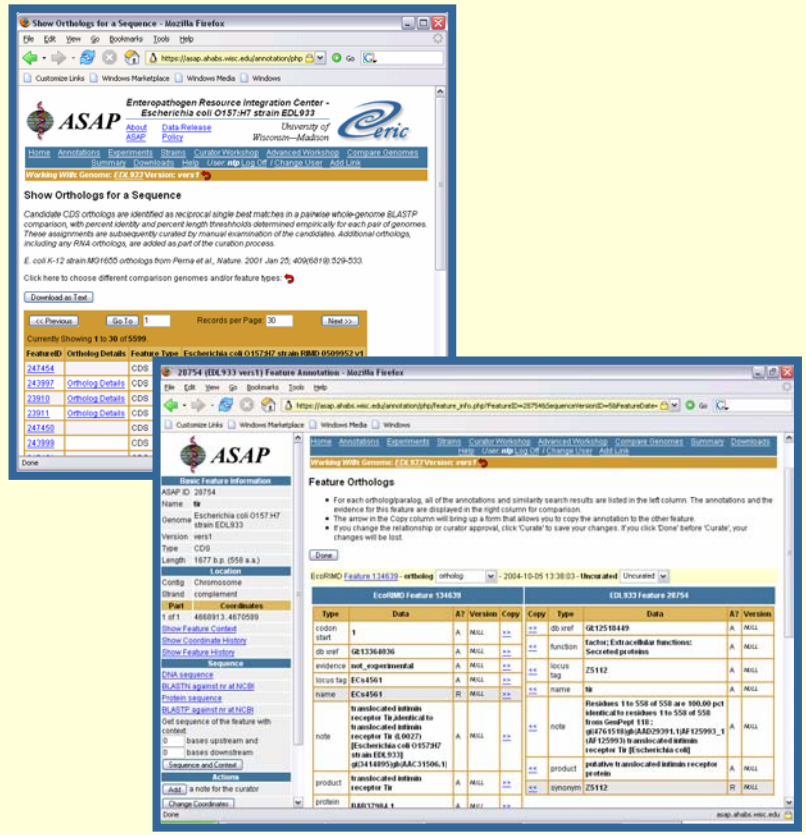


Enteropathogen Resource Integration Center

ERIC-BRC Annotation

Orthologs

- Semi-automated ortholog determination using filtered BLASTP reciprocal best hits
- Reviewed, corrected and augmented by curators
- Annotators/curators can transfer annotations from one ortholog to another
- Lists of shared and unique genes to query and download.



Enteropathogen Resource Integration Center

ERIC-BRC Annotation

Mauve multiple alignments, with display of annotated features